FACULTY OF
BUSINESS &
ECONOMICS

Melbourne Institute Working Paper Series

Working Paper No. 44/13

The Measurement of Cognitive Ability in Wave 12 of the HILDA Survey

Mark Wooden

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

The Measurement of Cognitive Ability in Wave 12 of the HILDA Survey*

Mark Wooden
Melbourne Institute of Applied Economic and Social Research
The University of Melbourne

Melbourne Institute Working Paper No. 44/13

ISSN 1328-4991 (Print)

ISSN 1447-5863 (Online)

ISBN 978-0-7340-4338-2

December 2013

* This paper was prepared as part of the management contract the Melbourne Institute has with the Australian Government Department of Social Services (DSS) for managing and administering the Household, Income and Labour Dynamics in Australia (HILDA) Survey. It uses a near final version of Release 12 of the HILDA Survey unit record data set, only available to members of the Melbourne Institute HILDA Survey Project team. As such, there may be small discrepancies between the numbers reported and those that can be derived from the final and public version of Release 12. The HILDA Survey is a project initiated and funded by the DSS and managed by the Melbourne Institute of Applied Economic and Social Research. The findings and views reported in this paper, however, are those of the authors and should not be attributed to either DSS or the Melbourne Institute. The data are available for research purposes under license. Details of how to obtain the data can be found at <http://melbourneinstitute.com/hilda/>.

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne
Victoria 3010 Australia
***Telephone* (03) 8344 2100**
***Fax* (03) 8344 2111**
***Email* melb-inst@unimelb.edu.au**
***WWW Address* http://www.melbourneinstitute.com**

Abstract

This paper provides a statistical overview of three new cognitive ability measures collected in wave 12 of the HILDA Survey: (i) Backwards Digit Span; (ii) the Symbol Digits Modalities Test; and (iii) a 25-item version of the National Adult Reading Test. The paper: analyses willingness to participate (as reflected in item response); examines the evidence on the performance of the measures and the extent to which scores were influenced by external circumstances, interview mode, and interviewers; reports summary descriptive data; and presents cursory evidence on the extent to which the measures are predictive of outcomes, and more specifically the hourly wage among working-age employees.

JEL classification: C81, I20, J24

Keywords: Cognitive ability, HILDA Survey, survey measurement, panel survey data, wages

Introduction

In a previous paper, Wooden et al. (2012) describe the identification, development and testing of suitable measures of cognitive ability for inclusion in wave 12 of the HILDA Survey. As reported on there, three measures were proposed for inclusion. That recommendation was subsequently approved by what was then the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (now Department of Social Services).

In this paper the data on cognitive ability collected in wave 12 are examined. Specifically, I: analyse willingness to participate (as reflected in item response); examine the evidence on the performance of the measures and the extent to which scores were influenced by external circumstances, interview mode, and interviewers; report summary descriptive data; and present cursory evidence on the extent to which the measures are predictive of outcomes, and more specifically the hourly wage among working-age employees. The paper concludes with a summary of issues that researchers should be aware of when using these data.

The Measures

The three measures of cognitive ability included in wave 12 of the HILDA Survey were: (i) Backwards Digit Span; (ii) Symbol Digits Modalities; and (iii) a 25-item version of the National Adult Reading Test. A paper representation of the script used to administer these tasks is provided in Appendix A. Note, however, that almost all interviews in wave 12 of the HILDA Survey were administered using computer-assisted methods[1], and hence there may be small differences between the paper representation provided here and the much longer and more complex script that is generated for computer delivery.

Backwards Digit Span (BDS)

Backwards Digit Span is a test of working memory span. It features in many traditional intelligence tests, including the Wechsler Adult Intelligence Scales. It involves interviewers slowly reading out successively longer strings of single-digit numbers and asking participants to repeat those strings in reverse order. Respondents are given two chances at each length or level. When the respondent gets one trial correct at a level, the first trial at the next level is administered. If the first trial is incorrect, the second trial is administered. If both responses at the same level are incorrect, the test is discontinued. The longest sequence administered is eight digits.

The format used in the HILDA Survey is taken directly from the Brief Test of Adult Cognition by Telephone (BTACT), which was developed for the second wave of the Midlife in the US (MIDUS) study (Lachman & Tun 2008). While this version was developed specifically for telephone administration, the test was originally intended for administration in a face-to-face setting.

Symbol Digits Modalities (SDM)[2]

The Symbol Digits Modalities test was originally developed as a screening measure for cerebral dysfunction, but has been widely used in broader settings as a general test for divided attention, visual scanning and motor speed (Strauss et al. 2006, p. 617). It involves participants matching symbols to numbers using a printed key. The score is the number of

[1] A total of 37 person interviews were conducted using pen-and-paper methods in wave 12.

[2] The test is protected by copyright, which is owned by Western Psychological Services (WPS). A license to use the test in wave 12 of the HILDA Survey was purchased from WPS.

items correctly matched within a 90 second time interval. The task can be administered either orally or in written form. For the HILDA Survey we adopted the latter approach. It is not suitable for telephone delivery (given the presence of the printed key that has to be handed to the respondent just prior to the task commencing) and so was not administered to any respondents interviewed by phone. Further details about the test, including a copy of the test answer sheet, can be found in the test manual (Smith 2007).

National Adult Reading Test, Short-Form (NART25)

The National Adult Reading Test (NART) is a reading test of 50 irregularly spelled words, listed roughly in order of difficulty, which is intended to provide an estimate of pre-morbid intelligence. The value of the test lies, in part, in the high correlation between reading ability and intelligence in the normal population, with numerous studies reporting moderate to high correlations between NART performance and measures of intellectual status (see Strauss et al. 2006). Indeed, scores on the NART are designed to predict scores on the WAIS-R intelligence test. Further details about the test rationale, as well as procedures for administration and scoring are provided in the test manual (Nelson 1982).

A major impediment to the inclusion of cognitive ability measures in the HILDA Survey is the time taken to complete them. As summarised in Wooden et al. (2012), for the wave 12 pre-test (or Dress Rehearsal) a total of five measures were included, with the total time taken averaging almost 18 minutes for any respondent who completed all five tasks. We were thus compelled to remove measures (the proposed category fluency and number series tasks were not included in the main survey instrument) and to consider how administration of the remaining tasks could be made more time effective. We thus used the data collected from the pre-test sample to estimate parameters from item response theory models to determine whether a word list half the length of the original NART could be administered that would retain measurement properties similar to the original. A 25-item version (NART25) was found to be highly reliable, as measured by Cronbach's alpha (0.89) (which compares with 0.92 for the full test), and highly correlated with the 50-item version (r=0.97).[3] The NART25 was thus administered in wave 12.

The NART was designed to be administered to persons aged 18 years or over, though in Australia seems to have been mainly used in studies of older populations (e.g., Kiely et al. 2011). In the HILDA Survey, however, we applied no additional age restrictions, and hence test participants include persons as young as 15 years.

An obvious weakness with the task is that it is only intended for persons who can read English. Indeed, a person who cannot read any English will be unable to undertake the task. It is thus not an appropriate measure of intelligence for non-English language speakers or for persons whose reading ability has been seriously compromised by injury or illness. Nevertheless, given we are also interested in reading ability in its own right, there are still good reasons to administer the task to non-native English language speakers in the HILDA Survey sample. That is, we anticipate that NART scores could also double as a measure of functional literacy.

The task involves participants being presented with a word card and instructed to read out loud each word. Interviewers then record correct pronunciations, with the total correct providing the score. Slight variations in pronunciations due to regional accents are acceptable. Again it is not easily delivered over the telephone, and hence was only administered to HILDA Survey respondents interviewed face-to-face.

[3] The Cronbach's alpha derived for the NART25 from the main wave 12 survey is also 0.89.

For the HILDA Survey, pronunciations that are regularly used in not just Australia (as reflected in the Macquarie Dictionary), but also in the UK and the US, were permitted. That said, only five of the words from the list of 25 were determined to have acceptable multiple pronunciations.

Administration of all tasks required additional time and resources to be devoted to interviewer training. This was particularly important with respect to NART25, the administration of which places considerable burden on interviewers to identify correct and incorrect pronunciations. Previous research on the NART indicates that inter-rater reliability is actually very high – typically above 0.88 (Strauss et al. 2006, p. 196). However, it is possible that the test administrators used in those other studies are very different to the average interviewer employed on the HILDA Survey.

Task Participation

Summary data describing the number of responding sample members in wave 12 who were invited to complete each task, the numbers agreeing, and the number of valid scores that were generated are presented in Table 1.

A total of 17,476 persons were interviewed in wave 12 of the HILDA Survey, but five of these persons terminated the interview prior to the section on cognitive ability tasks, which was placed very close to the end of the interview. A total of 17,471 persons were thus invited to participate in the BDS task. For the other two tasks the initial sample is smaller given persons interviewed by telephone were not given the option of participating; 1383 persons were interviewed by telephone in wave 12 (7.9% of the sample).[4]

Respondents then had the option of opting out from each task. In addition, it was not possible to administer some tasks to some respondents given an inability to understand the instructions. The numbers of non-participants, however, were relatively small, and hence task participation rates were very high; 95% to 96%.

Table 1: Summary of task participation

Sub-sample	*BDS*	*SDM*	*NART25*
Invited to participate	17471	16091	16090
Unable to understand instructions	148	175	158
Refused	569	630	626
Number of task participants (Participation rate I, %)	*16754 (95.9)*	*15286 (95.0)*	*15306 (95.1)*
Task not completed in full	240	112	234
Assisted by someone else in household	73	50	66
Number of persons with valid scores (Participation rate II, %)	*16446 (94.1)*	*15125 (94.0)*	*15007 (93.3)*

[4] A further respondent commenced the interview face-to-face, but completed it at a later date by telephone.

The number of valid scores obtained is slightly lower again due to some persons not completing the task in full (because of interruptions or because of an unwillingness to continue) or because the interviewer reported that the respondent received outside assistance. This reduced the measured participation rates by a further one to two percentage points, to 93 to 94%.

While task participation rates are very high, non-participation may still be problematic if the relatively small group of non-participants is markedly different from the participating group. In particular, we might expect that persons with cognitive deficits to be both less able and less willing to complete the tasks. These expectations appear to be confirmed in Table 2, which reports task participation rates by selected respondent characteristics: sex, age, educational attainment, origin (or more specifically whether born in Australia or overseas, and if the former, whether an Indigenous Australian, and if the latter, whether English was their first language), and various self-assessed measures of ability.

Thus we can see that participation rates are lowest among sample members who: are elderly; did not complete high school; have work-limiting long-term health conditions; are Indigenous Australians; did not speak English as their first language; and generally regard their skills in either speaking English, reading English or mathematics as poor compared to the average Australian. Clearly non-random selection effects are at work. The effects also seem to be more acute with respect to the NART25, where the ability to recognise at least some English language words is a precondition of participation.

However, while the differences reported in Table 2 are, with the exception of sex, always statistically significant, many could be argued to be relatively modest in size. That is, task participation rates are still very high for most groups. For example, while participation rates are significantly lower among persons aged 75 years or older, almost 88% of this group still participated in the BDS and NART25 tasks, while a smaller, but still sizeable fraction (82%) completed the SDM task. Further, the groups where participation rates are very low, such as persons with severe disabilities and persons who speak or read English poorly, represent a relatively small fraction of the total sample.

Finally, I also examined whether the likelihood of task participation was related to a respondent's HILDA Survey participation history or to the interview situation. No evidence of any strong statistical association between the number of times a respondent had previously responded to the survey and rates of task participation could be found. Similarly, differences in participation rates between members of the original sample and members of the top-up sample introduced in wave 11 were very modest, and most likely due to the higher incidence of recently arrived immigrants in the latter group. I thus have chosen not to report results for these variables in Table 2.

Table 2, however, does report participation rates differentiated by both survey mode and survey location. Unsurprisingly, interviews that are conducted outside the home (e.g., on the doorstep) are less conducive to task completion. The same is true of interviews conducted in other places (e.g., cafes, public parks). Nevertheless, it is again true that even in these less favourable circumstances, task participation rates are very high. And again it must be remembered that the vast majority of interviews are conducted inside respondents' home (or if a telephone interview, with respondents who are at home).

The evidence also suggests that the presence of another adult while the interview was being conducted had an inhibiting effect on task participation, with task participation rates being three to five percentage points lower when someone else was present.

Table 2: Participation rates by selected characteristics

Characteristic	*% distribution*	*Participation rate (%)*[a]		
		BDS	*SDM*	*NART25*
Sex				
Men	47.3	94.2	94.0	93.0
Women	52.7	94.1	94.0	93.4
Age group				
15-24	18.7	96.4	96.9	94.6
25-34	17.0	95.9	96.3	94.6
35-44	16.8	95.5	95.2	94.7
45-54	16.8	93.8	94.6	93.8
55-64	13.8	93.1	93.4	92.2
65-74	9.8	91.0	91.8	91.0
75+	7.1	87.9	82.2	87.7
Highest education level				
Less than Year 12	30.4	91.1	90.1	87.9
Year 12	16.2	94.8	95.3	95.2
Trade certificate (III / IV)	21.4	95.5	95.4	94.6
Diploma	8.6	95.7	96.4	96.3
Bachelor's degree	13.7	95.8	96.5	96.4
Higher degree / qualification	9.7	96.0	95.9	97.3
Long-term health conditions				
None	72.8	95.5	95.9	95.0
Mild (does not limit work at all)	8.5	92.2	92.2	92.1
Moderate	16.2	91.3	90.3	89.8
Severe (cannot work at all)	2.5	78.9	70.1	72.3
Origin				
Australia-born: Indigenous	2.8	93.5	93.2	87.4
Australia-born: Other	75.1	95.2	95.2	94.9
O/S: English first language	11.1	95.8	95.2	96.2
O/S: English not first lang.	11.0	85.2	85.0	80.8
Self-assessed English speaking ability				
Speaks only English	88.0	95.2	95.1	94.9
Very well	7.2	94.0	93.4	92.7
Well	3.2	84.0	85.0	77.9
Not well or Not at all	1.6	57.2	58.6	41.4
Self-assessed relative English reading skills (on 11-point scale)				
0-2	1.6	58.7	53.2	28.1
3	1.1	83.8	82.1	66.3
4	2.0	88.3	88.2	74.7
5	9.4	91.3	91.8	87.3
6-7	18.0	93.9	94.2	92.7
8-10	67.9	95.9	95.8	97.1

Table 2 (cont'd)

Characteristic	*% distribution*	*Participation rate (%)*[a]		
		BDS	*SDM*	*NART25*
Self-assessed relative mathematics skills (on 11-point scale)				
0-2	2.1	78.7	74.4	66.2
3	2.6	88.7	90.1	83.5
4	4.7	89.3	91.2	87.5
5	16.0	93.0	92.3	91.4
6-7	30.4	94.9	95.0	94.8
8-10	44.5	95.8	95.6	95.6
Interview mode and location				
Face-to-face: in home	87.9	94.3	94.1	93.5
Face-to-face: outside home	1.9	91.9	91.9	88.0
Face-to-face: at a workplace	0.7	94.8	95.7	96.5
Face-to-face: other place	1.6	91.4	89.6	86.1
Phone: at home	6.9	93.2	-	-
Phone: other place	1.0	89.9	-	-
Presence of other adults during interview				
Others present	38.1	91.6	91.9	90.3
Alone	61.9	95.8	95.4	95.2

Notes: a The participation rate used here defines a sample as a participant only if their participation resulted in a valid score on the task.

NS denotes not statistically significant at the 0.10 level or better.

Multivariate Analysis

All of the conclusions drawn above are based on simple cross-tabulated data. I thus next tested whether these conclusions were robust within a multivariate framework where all of the observed characteristics are controlled for simultaneously.

Since the outcome variables – identifying participation in each task – are all dichotomous variables, logistic regression was employed. Results (estimated logit coefficients plus standard errors) are presented in Table B.1 (in an Appendix). For each outcome variable, I present results from two specifications; one which controls for interviewer effects (specification B) and the other which does not (specification A).

The results broadly support the earlier conclusions, with *relatively* low rates of task participation among persons with work-limiting long-term health conditions, with poor English language skills, and who did not complete high school, and with interview conditions (the presence of others and the location of the interview) also playing a role. There are, however, some interesting differences and nuances.

First, the negative effect of age only kicks in at a very old age, and even then there is no diminished rate of participation in the NART25 by age. This is consistent with the notion that participants are more likely to participate in these tasks if they believe they will fare well – as we will see later, scores on the NART25 rise with age.

Second, the effects of severe long-term health conditions on task participation are very large, with the implied relative odds of participation (or odds ratios) varying between just 0.14 and 0.22.

Third, the lower rate of task participation among Indigenous Australians is only significant (and pronounced) for the NART25, suggesting that it is discomfort with English literacy skills that was the main deterrent to Indigenous respondents.

Fourth, while controlling for interviewer effects does not fundamentally alter the pattern of estimated coefficients, the magnitude of some estimates change considerably. As an aside, note that when controlling for interviewer effects we may also be capturing the influence of location-specific factors on task participation, since the face-to-face interviewers are typically assigned to specific neighbourhoods.

Cognitive Ability Task Scores

Descriptive Statistics

Summary descriptive statistics for the scores on each of the three cognitive ability tasks are presented in Table 3. The scores for each are derived as follows:

- Achievement on the BDS task is reflected in the highest number of digits correctly recalled and repeated in reverse order and is usually scored as the length of the longest correct sequence, and thus can take the values 0 and 2 to 8. It is recommended here, however, that a value of 1 be subtracted from all positive scores, resulting in a variable with a range of 0 to 7.
- As described earlier, the NART25 requires correctly pronouncing 25 irregularly spelt words from a card (as judged by trained interviewers), and hence the score is just the number of correctly pronounced words. Scores obviously can range from 0 to 25.
- Also as described earlier, the SDM task involves matching numbers to symbols using a keycard. The score is simply the number of correct matches achieved within a 90-second time frame.

As the statistics in Table 3 indicate, all of the tests have properties that suggest distributions that are close to normal. BDS scores have a distribution that is somewhat skewed to the right (top half of the distribution), while scores on both NART and SDM are slightly skewed to the left (the bottom half).

Further, while all test scores are positively correlated with each other, the size of that correlation is not so large to suggest that any of the tasks are redundant. This can be seen in Table 4, where the simple correlations are in the range of .2 to .4. The smallest correlation is between SDM and NART (.21), suggesting these are the least substitutable. However, once age is held constant, this correlation almost doubles in size.

Table 3: Cognitive ability task scores – summary statistics

Statistic	*BDS*	*SDM*	*NART25*
Mean	3.91	49.18	13.53
Standard deviation	1.44	13.11	5.44
Percentile distribution			
10th	2	32	6
25th	3	41	10
50th (median)	4	50	14
75th	5	58	18
90th	6	65	20
Minimum	0	0	0
Maximum	7	110	25
Skewness	.37	-.33	-.25
Kurtosis	-.43	.53	-.51
N	16446	15116	15007

Table 4: Correlations between cognitive ability task scores

	BDS	*SDM*	*NART25*
BDS	1	.338 (.334) [.301]	.374 (.402) [.367]
SDM		1	.210 (.388) [.326]
NART25			1

Note: Figures in round parentheses are age-adjusted correlations, while figures in square parentheses adjust for age, the presence of moderate and severe long-term health conditions, not completing high school, and relatively poor English language speaking ability.

I next checked whether cognitive ability scores vary in expected ways with both age and education level. Thus in Table 5, data on mean scores cross-classified by both age group and a crude (binary) indicator of educational attainment are presented. As expected, scores on the three tasks are positively correlated with educational achievement, with persons with university qualifications averaging the highest scores within all age groups (and for all tests) and persons who did not complete Year 12 (or its equivalent) averaging the lowest scores. Also evident is that persons with trade qualifications (Certificate level III or IV) invariably score worse than persons who completed Year 12 but did not go on to obtain any post-school qualification.[5]

It can also be seen that the scores exhibit very clear associations with age. BDS and SDM both tend to decline with age. This is especially pronounced in the SDM task. The decline in performance on the BDS task, on the other hand, only starts to become pronounced once people enter retirement years (that is, from age 65). In contrast, scores on the NART25 typically rise with age. These divergent patterns are entirely in line with expectations and reflect the types of ability / intelligence – and specifically the distinction between fluid and crystallized intelligence (Cattell 1971) – that are intended to be captured by these measures. Fluid intelligence is the capacity to think logically and solve problems independent of acquired knowledge. Crystallized intelligence is the ability to use skills, knowledge, and experience. The former is expected to be closer to innate endowments, and will deteriorate at older ages. In contrast, the latter improves with experience and hence with age. The BDS and SDM tasks were expected to be much more closely associated with fluid intelligence, while NART25 is expected to tap into one aspect of crystallized intelligence.

[5] As per the decision framework used by the Australian Bureau of Statistics (see explanatory note 42 in *Education and Work, Australia, May 2012*, ABS cat. no. 6227.0), persons with a Certificate Level I or II are coded as having an educational level no higher than completing Year 10 of high school, and so are all included in the lowest education category reported in Table 5.

Table 5: Mean cognitive ability scores by age group and highest education level

	Education level					
Task / Age group	*Less than Year 12*	*Year 12*	*Certificate III / IV*	*Diploma*	*Degree*	*Higher qual.*
BDS						
<18	3.80	3.91	*	-	*	-
18-24	3.43	4.12	3.88	3.86	4.35	4.72
25-34	3.63	3.96	3.82	3.95	4.52	4.46
35-44	3.60	4.02	3.78	4.17	4.51	4.51
45-54	3.62	4.28	3.76	4.17	4.36	4.38
55-64	3.48	3.94	3.72	4.12	4.40	4.49
65-74	3.48	3.87	3.50	3.86	4.15	4.26
75+	3.12	3.68	3.17	3.66	3.90	3.80
SDM						
<18	53.97	60.90	*	-	*	-
18-24	49.17	57.55	54.42	55.86	61.26	57.89
25-34	49.60	55.70	53.70	57.07	59.06	59.07
35-44	48.35	53.12	50.67	54.66	56.27	57.46
45-54	46.55	50.08	47.17	51.08	52.67	53.10
55-64	42.87	45.63	43.76	47.73	48.58	49.07
65-74	36.03	40.25	37.01	40.53	42.37	44.58
75+	27.24	30.59	28.11	33.63	33.08	37.46
NART25						
<18	9.88	11.63	*	-	*	-
18-24	8.97	12.76	11.29	11.53	15.05	15.43
25-34	9.95	12.90	11.43	12.79	14.81	14.69
35-44	10.60	13.39	12.30	13.98	16.30	16.88
45-54	11.56	15.00	12.68	15.07	17.06	17.99
55-64	12.48	15.19	13.18	16.05	17.92	19.04
65-74	12.58	15.65	13.95	16.57	18.15	19.74
75+	12.41	16.52	13.60	17.44	18.47	20.55

Note: * Cell includes very few cases (<20) and hence mean score for these cells not reported.

Also as expected, task scores vary with a sample member's background, and more specifically with whether English was their first language and with their self-assessed fluency in English. This should be evident from the patterns in mean scores reported in Table 6. Not surprisingly, the differences are most marked with the NART25, given this is only a valid test of (verbal) intelligence if the respondent is a native English speaker. Native English speakers fared best on this task, though interestingly native English-speaking migrants fared noticeably better than their Australian-born counterparts. And scores decline rapidly with the level of proficiency in English speaking.

For the other two tasks, differences by origin, conditional on speaking English at least "very well", are not pronounced. However, scores again drop markedly for persons with lesser levels of self-assessed English language ability, despite the fact that performance on the tasks involved did not depend on good English language skills (though some functionality in the English language will have been necessary in order to comprehend the task instructions).

Table 6: Mean cognitive ability task scores by origin and self-assessed English speaking ability

Task / English language speaking use and ability	*Origin* Australia-born: Indigenous	Australia-born: Other	Overseas-born: English first language	Overseas-born: English not first language
BDS				
Only speaks English	3.43	3.96	4.00	3.77
Speaks English				
Very well	*	3.89	4.11	3.91
Well	*	3.62	2.90	3.41
Not well / Not at all	*	*	*	3.08
SDM				
Only speaks English	47.84	49.67	47.58	47.61
Speaks English				
Very well	*	53.36	50.02	50.42
Well	*	46.33	42.24	44.10
Not well / Not at all	*	*	*	38.40
NART25				
Only speaks English	9.97	13.83	15.37	13.29
Speaks English				
Very well	*	13.58	12.74	11.49
Well	*	10.52	5.36	7.53
Not well / Not at all	*	*	*	4.52

Note: * Cell includes very few cases (<20) and hence mean score for these cells not reported.

Influence of External Circumstances

Ideally, each of the three tasks should have been administered in a quiet environment with no one else present, but in surveys conducted in private homes this is often not possible. Indeed, around 24% of all interviews were conducted with someone else present at the time the cognitive ability tasks were conducted.[6] Moreover, and as shown in Table 6, the presence of others during the interview was associated with a lower score on all tests, suggesting the presence of others may often have been a distraction. The magnitude of these differences, however, while always statistically significant, is only of any size for the NART25. Further, it does not necessarily follow that these differences imply the presence of others had a causal influence on task performance. For example, the NART25 is the only test where the presence of children is associated with markedly lower scores when compared with being interviewed alone. But scores on the NART25 tend to rise with age, while persons with young children are often relatively young themselves. It thus may be that part of the observed difference in scores is nothing more than an age effect – that is, persons in households where other persons

[6] The proportion of interviews where someone else is recorded as being present during the interview is much larger– 38% of all interviews are recorded as having another adult present, and of course there will be other interviews where children (but not any other adult) are present. The lower proportion reported here simply reflects the protocol that interviewers do their best to ensure that others leave the room when it is time to administer the cognitive ability tasks.

are present (and especially children) are relatively young and thus likely to score not as highly on the NART25.

More illuminating may be data collected from interviewers on whether performance on each task was adversely affected by any significant distraction or disturbance. Such distractions, of course, could be due to a wide range of events or circumstances, which might occur with or without the presence of others in the home (for example, the respondent answering a phone call). The summary data reported in Table 7 suggest that such disturbances were relatively uncommon, affecting between 3% and 8% of respondents, depending on the task. More importantly, the mean scores on both the BDS and SDM tasks were not significantly lower among persons reported as being affected by a disturbance. Only performance on the NART25 seems to have been seriously affected by disturbances, and in this case the proportion affected was relatively small (just 3.1% of task participants).

Table 6: Mean cognitive ability task scores by presence of others at time of interview

	BDS	*SDM*	*NART25*
No one else present	3.95	49.38	13.82
Others present	3.81	48.66	12.76
(T-test)	(5.89**)	(2.99**)	(10.66**)
Type of person present			
Other sample member	3.80	48.19	12.85
Child / children under 15	3.91	52.99	11.86
Non-household member	3.86	48.55	12.34

Note: ** indicates that the difference in means is significantly different from zero at the 99% confidence level.

Table 7: Mean cognitive ability task scores by whether interview affected by distractions or disturbances

	BDS	*SDM*	*NART25*
Incidence of significant distractions or disturbances (%)	7.6	4.3	3.1
Mean test score			
Affected by significant distraction or disturbance	3.92	49.15	12.84
Not affected	3.91	49.19	13.55
(T-test)	(0.05)	(0.07)	(2.62**)

Note: ** indicates that the difference in means is significantly different from zero at the 99% confidence level.

Mode and Location Effects

As noted earlier, while the large majority of interviews are conducted face-to-face within the respondent's home, there are departures from this norm, raising the possibility that scores might be sensitive to both mode and location. Most notably almost 8% of interviews are conducted by telephone, and while telephone respondents were not administered either the

SDM or NART25 tasks, they were asked to complete the BDS task. The potential problem, here, however, is there is greater scope for participants to write numbers down as they are read out, which would lead to superior scores on this tasks.

As can be seen from Table 8, the telephone respondents do score better on average, but the difference, while statistically significant, is quite small (especially if we only focus on persons interviewed while at home). However, if some telephone respondents had been writing down the numbers as they were read out to them, we would expect the differences to be entirely concentrated at the upper end of the distribution; that is, we would expect many more telephone respondents to be able to correctly repeat in reverse order the longest string (8-digits). These expectations are confirmed. While only 5.1% of participants interviewed face-to-face achieved the top possible score, more than double (11.1%) that proportion of telephone interviewees successfully completed the longest sequence. Unless selection into telephone interviews is strongly correlated with ability, this large difference, combined with the relatively small difference in mean scores, is strongly suggestive of "cheating" by a small subset of the telephone respondents.

Table 8: Mean cognitive ability task scores by interview mode and location

Interview mode / Location	*BDS*	*SDM*	*NART25*
Face-to-face: In home	3.90	49.16	13.56
Face-to-face: Outside home	3.89	46.57	12.36
Face-to-face: Respondent's workplace	4.20	53.35	14.54
Face-to-face: Other place	3.97	52.14	12.78
Phone: At home	4.03	-	-
Phone: Other place	4.49	-	-
Total	3.91	49.18	13.53

Otherwise the mean BDS scores do not seem to vary much with interview location, consistent with the notion that the BDS task is relatively simple to administer and only requires a place where the respondent can concentrate for a short period. The notable exception is persons interviewed in their workplace, whose mean score is notably above the sample mean. This almost certainly reflects the characteristics of persons who are likely to opt for an interview at their workplace – mostly managers and professionals – rather than any inherent feature of the location.

Very differently, mean scores on both the SDM and NART25 tasks are significantly lower among persons interviewed outside the home (i.e., at the door or in the yard) when compared with persons interviewed inside their home. This is consistent with the expectation that tasks administered outdoors will generally be less conducive to high scores than interviews administered indoors.

Interviewer Effects

An issue of concern, especially for the NART25, is the possibility that test scores are correlated with interviewer ability. A simple test of interviewer ratings conducted during interviewer training prior to the wave 12 pre-test revealed relatively poor performance on the full NART, with interviewers averaging two incorrectly scored items each, but with a number of notable outliers (the maximum number of errors was 14) all of whom were not native English speakers.

Identifying interviewer effects, however, is complicated by the fact that respondents are not randomly distributed across interviewers. Interviewers are allocated workloads that are geographically clustered and hence mean ability scores when averaged across interviewers will be affected by any factor that is both correlated with location and ability (e.g., socio-economic status). Thus we do expect systematic variation by interviewer, but because of factors that are correlated with location, rather than because of systematic biases caused by the interviewer.

However, we can take advantage of the strong likelihood that the scope for interviewer effects is considerably reduced in some of the tasks. Most obviously, the Symbol Digits Modalities task is largely completed by the respondent with minimal interviewer involvement. The interviewer's only responsibilities are to read out the instructions and to monitor compliance with the 90 seconds time limit. We would expect very little of the observed variation between interviewers in SDM scores to be due to variance in interviewer ability or testing methods.

In Table 9, the results from a simple one-way analysis of the variance in cognitive ability scores where the independent variable is interviewer identity are reported. The data used are restricted to test scores collected in face-to-face interviews and only from interviewers that conducted a minimum of 10 interviews (n=144). As can be seen, relatively little of the variance is due to differences in scores between interviewers. It is less than 5% for the BDS task, less than 7% for SDM task, and just under 12% for the NART25. As expected, interviewer effects are largest for the NART25, which might suggest that inter-interviewer reliability in scoring is not as good as desired. That said, the proportion of variance explained by between-interviewer effects is much smaller than recorded in the pre-test (17.9%), presumably reflecting the additional efforts put into interviewer training. Further, it also seems likely that NART scores will be much more affected by factors associated with location (such as age and socio-economic status) than the other two cognitive ability tasks.

Table 9: Analysis of variance in cognitive ability scores by interviewer

% of sum of squared residuals due to:	*BDS*	*SDM*	*NART25*
Within interviewer effects	95.3	93.3	88.3
Between interviewer effects	4.7	6.7	11.7
F	5.16**	7.54**	13.72**

Notes: Sample restricted to face-to-face interviews conducted by interviewers that conducted at least 10 person interviews.

** indicates that the difference is significantly different from zero at the 99% confidence level.

Do Variations in Task Scores Simply Reflect Variations in Motivation?

A final measurement concern is the extent to which scores reflect differences in respondent motivation and effort. While these tasks are intended to test respondents' abilities, the tasks are not being administered under test conditions where respondents have incentives to perform as best they can. To test the extent to which this is so, correlations between task scores and a crude measure of achievement motivation (derived from data collected in the self-completion questionnaire) are examined. On the assumption that task motivation is highly correlated with underlying achievement motivation, sizeable correlations between a measure of achievement motivation and cognitive ability task scores might indicate that the ability tasks are contaminated by differences in the motivation and ability to perform well (though that said, it should be recognised that achievement motivation may itself be a function of ability).

The achievement motivation measures used were multi-item scales based on the work of Lang and Fries (2006). The intent was to measure the twin constructs "hopes for success" and "fear of failure". Scale construction began with nine items scored on a 7-point disagree / agree scale.[7] As expected, principal components analysis extracted two factors with eigen values greater than one from the responses to these nine items, with the items all loading highly on the expected factor (loadings of .62 or higher) and lowly on the other factor (no loadings greater than .18). Strong support thus exists for the construction of two independent scales, one measuring hopes for success (and comprising four items) and the other measuring fear of failure (and comprising five items). Simple additive scales also prove to be highly reliable, as measured by Cronbach's alpha (.84 for the fear of failure scale and .74 for the hopes for success scale). Further, the correlation between the two is relatively modest – just -.25.

Correlations between these motivation scales and the cognitive ability task scores are presented in Table 10. As can be seen, ability scores are correlated with achievement motivation, but the two dimensions have opposite effects: persons with strong desires to achieve score better on the tasks while those who fear failure tend to score worse. While no causal interpretation can be placed on simple correlations, these findings are more consistent with reverse causation. If motivation promotes better task scores then we would have expected both hopes for success and fear of failure to be positively correlated with scores. In contrast, if the causation runs from ability to motivation, then high ability persons would typically be expected to both favour situations where they are challenged and tested and to be relatively less likely to fear failure. The latter interpretation is the one that is consistent with the data.

Table 10: Correlations between achievement motivation sub-scales and cognitive ability task scores

Cognitive ability task	*Hopes for success*		*Fear of failure*	
	r	*Partial r**	*r*	*Partial r**
BDS	.070	.046	-.075	-.060
SDM	.077	.056	-.053	-.091
NART25	.082	.048	-.107	-.059

Note: * Controlling for sex, age, the presence of moderate and severe long-term health conditions, not completing high school, and relatively poor English language speaking ability.

[7] The measure proposed by Lang and Fries (2006) included 10 items. We also amended the wording of a number of the original items to be more appropriate for an instrument administered in English to an Australian population.

Finally, the magnitudes of these correlations are very small, and typically decline further once other likely covariates are held constant. So even if there is a direct causal effect, it may be relatively unimportant.

Multivariate Analysis

But are the conclusions reached in the preceding discussion robust once the various correlates of task performance are simultaneously controlled for? I thus again estimated regression models that control for a range of characteristics describing both the respondents and their interview situations. Least squares regression methods were used, which in theory require the dependent variables to be continuous. This is not true for any of the three outcomes. All are discrete variables, and in the case of both the BDS and NART25, scores are constrained to lie within pre-determined ranges. The linear regressions estimations reported on here, therefore, should be treated as rough approximations, and not too much emphasis should be placed on the precise magnitudes of the estimated coefficients.

The detailed results are again presented in Appendix B (see Table B2), and confirm the patterns observed in the descriptive data with respect to age, health conditions, origin, English speaking ability and education.

The estimates also confirm that the presence of someone else while the task was being conducted had a negative influence on the performance on all three tasks, though again the size of these effects seem to be quite small. And again we can see that interviewer observations about task performance being adversely affected by significant distractions or disturbances only appear to be reflected in significantly lower scores on the SDM task.

Interview mode and location effects again are found to be small or absent, with telephone interviews associated with a slightly better performance, on average, on the BDS (but bear in mind the caveat that these effects are likely to be concentrated on a relatively small group of "cheats" for whom the differential will be much larger), and with interviews conducted outside or at the door having a significantly deleterious effect only on NART25 scores.

Finally, the two achievement motivation scales exhibit significant associations with all three outcomes, with the pattern of signs more consistent with the notion that motivation is a function of ability rather than the contrasting hypothesis that motivation directly affects task performance.

Cognitive Ability Measures and Real World Outcomes: The Case of Labour Earnings

The final feature of this paper is the consideration of the importance of these proxy measures of cognitive ability for real world outcomes, with the analysis reported below focussing on just one such outcome – hourly wages.

Cross-sectional wage equations are estimated, which take the form:

$$\ln W_i = X_i'\beta + Z_i'\gamma + \varepsilon_i$$

where the dependent variable is the log of the hourly wage (W), constructed here as the gross estimated usual weekly labour earnings divided by usual weekly hours of work[8], X is a vector of individual characteristics assumed to be related to earnings, and Z includes the measures of ability.

Again the estimation approach used is very simple – linear least squares regression. It would, for instance, be fairly straightforward to pool observations across all 12 survey waves and estimate panel data models, but which still allow for time-invariant regressors (e.g., random effects models). Additionally, no attempt has been made here to account for the potential selection bias arising from the fact that wages are only observable for persons in employment. The conventional approach to this problem is to estimate a selection model which simultaneously estimates an equation for the probability of employment. This, however, is not straightforward since it can often be difficult to identify variables that will satisfy the necessary exclusion restrictions. The results reported on here thus do not necessarily generalise to the wider population.

For purposes of this estimation, the sample is restricted to employees (where a person who is an employee of their own business is defined to be self-employed and hence out of scope) aged between 15 and 64 years. Any employees who are also full-time students at the time of interview are also excluded. I further excluded any cases where the value on the dependent variable seemed extreme. These boundaries, which admittedly are somewhat arbitrary, are \$8 and \$300 per hour. This resulted in the loss of 65 and 4 cases respectively from the sample, as just defined.[9]

The list of variables for inclusion in the X vector is potentially quite long given the richness of the HILDA Survey data. In the most detailed specification, controls are included for:

- sex;
- age (five-year age categories);
- marital status;
- the presence of a long-term health condition differentiated by whether it is work limiting or not;

[8] In the HILDA Survey the questions on hours of work specifically request respondent to include all hours of work, including both paid and unpaid overtime. Further, if queried, interviewers are instructed to advise respondents that this include work performed both in the home and at home, and that time "on call" is not considered usual work hours.

[9] At the time wave 12 of the HILDA Survey was conducted the Federal Minimum Wage (FMW) was \$15.96 per hour. The presence of employees reporting much lower hourly rates of pay in the HILDA Survey data reflects: (i) the presence of substantially lower minima for junior employees, apprentices / trainees and workers with a disability; (ii) the construction of hourly pay here using all reported working hours and not the standard weekly hours as specified in awards or agreements; (iii) non-compliance with award regulations; and (iv) reporting and measurement errors.

- origin (with three dummy variables included identifying whether the respondents is an Indigenous Australian, born overseas with English as their first language, and born overseas but with English not their first language);
- English language speaking ability (dummies which separate persons who speak a language other than English at home into three categories based on their self-assessed ability to speak in English);
- education (seven dummies identifying different levels of attainment);
- length of tenure (years) with the current employer (and its square);
- years of experience in the current occupation (and its square);
- union membership;
- contractual employment status (with two dummies identifying casual and fixed-term contract employment, respectively);
- employment through a labour-hire firm;
- occupation (dummies identifying 47 sub-major occupation groups are included);[10]
- industry of employer (dummies identifying 81 industry subdivisions are included);[11]
- sector of employer (with dummies included for public sector and private non-commercial);
- size of employer (with dummies included to identify firms with less than 20 employees and firms with 500 or more employees); and
- the two achievement motivation scales described earlier (which might be thought of as measures of aspects of non-cognitive ability).

The estimated coefficients on the three ability variables from a series of different specifications that include progressively more controls are reported in Table 11. Also reported is the impact of the inclusion of these three variables on the adjusted R^2 from the inclusion of the three cognitive ability variables, which provides a measure of the proportion of variance in the dependent variable that is explained, or accounted for, by these three explanatory variables.

In row 1 of this table are results from an equation that only includes the three cognitive ability variables. Together they account for just over 9% of the variation in hourly wages, but with the SDM test not significant. Further, the magnitude of the estimated coefficient on the BDS variable is very small, with a one standard deviation variation in the BDS score contributing to only a 1.4% difference in wages. In contrast, a one standard deviation variation in the NART25 score is associated with a 13.6% wage differential.

Once we account for demographic variables, such as age, sex and origin (row 2), the coefficient on the SDM task score increases in magnitude and becomes statistically significant, while the coefficient on the BDS task score becomes insignificant. These changes appear to be mainly driven by accounting for variations across age groups in the different task scores. But most importantly, the additional variance in wages explained by the cognitive ability variables declines to just 5.2%.

[10] Occupations are classified using the Australian and New Zealand Standard Classification of Occupations, 2006. Due to sample sizes, two groups had to be merged for this analysis.

[11] Industries are classified using the Australian and New Zealand Standard Industrial Classification, 2006. There are 86 industry subdivisions in this classification, but due to small sample sizes, the industry subdivisions within the broader Agriculture, forestry and fishing division have been combined to create the one industry category.

Table 11: Coefficients on cognitive ability task scores in hourly wage equations (standard errors in parentheses)

Dependent variable = Ln of usual gross weekly earnings divided by usual weekly hours of work
Estimation method = ordinary least squares

Control variables	*BDS*	*SDM*	*NART25*	*Adj.* R^2	*Change in* R^2	*N*
(1) None	.010** (.004)	.001 (.001)	.025** (.001)	.093		6911
(2) Demographics	.006 (.004)	.004** (.001)	.018** (.001)	.243	.052	6900
(3) Demographics + Education	.004 (.003)	.003** (.000)	.008** (.001)	.318	.013	6898
(4) Demographics + Education + Work experience	.005 (.003)	.003** (.000)	.007** (.001)	.333	.013	6896
(5) Demographics + Education + Work experience + Job / Firm characteristics	.002 (.003)	.002** (.000)	.005** (.001)	.471	.005	6499
(6) Demographics + Education + Work experience + Job / Firm characteristics + Motivation	.002 (.003)	.002** (.000)	.005** (.001)	.473	.005	6400
Sample						
(7) Females	-.001 (.004)	.002** (.001)	.004** (.001)	.425	.004	3174
(8) Males	.006 (.005)	.002** (.001)	.005** (.001)	.500	.004	3226
(9) Young (<30 years)	-.001 (.006)	.002* (.001)	.005** (.002)	.404	.005	1767
(10) Prime-age (30-49 years)	.005 (.005)	.003** (.001)	.005** (.002)	.450	.007	3021
(11) Mature-age (50-64)	.004 (.007)	.003* (.001)	.003 (.002)	.411	.003	1612
(12) Did not complete Year 12	-.007 (.007)	.002 (.001)	.005** (.002)	.406	.004	1016
(13) Completed Year 12 or a Level III / IV Certificate	.006 (.005)	.002** (.001)	.003* (.002)	.390	.005	2607
(14) University educated (Diploma or Degree)	.003 (.003)	.002** (.001)	.005** (.002)	.401	.005	2777

Notes:

1. All samples restricted to employees (and exclude any owner managers of incorporated enterprises) aged 15 to 64. Also excluded are full-time students, and any cases with outlying values on the
2. * and ** denote statistical significance at the 95% and 99% confidence levels, respectively.

Not surprisingly, this proportion declines even further with the inclusion of controls for education (row 3), given as we have already seen, cognitive ability and educational attainment are highly correlated. The cognitive ability measures now explain just 1.3% of the variance, and the magnitude of the estimated coefficients, especially on NART25, become even smaller. Inclusion of yet further controls (rows 4 through 6) reduce further the estimated contribution of the ability variables, though the additional decline is very small.

I also examined whether the effect of cognitive ability on wages might vary with the sex, age and education level of the workers. These results are reported in rows (7) through (14), with each specification including the complete set of controls. As should be apparent, the estimated wage effects of cognitive ability seem to be relatively robust to sample definition and selection, though statistical significance may decline (which mainly reflects use of smaller samples). There is, for example, no evidence that returns to ability vary with gender, as was found in an analysis of German panel data (Heineck and Anger 2010).

As a final check, I also tested for non-linearities in the relationship between cognitive ability and wages by replacing the scores with categorical valuables that sorted respondents according to where their scores placed them on the distribution (using the entire responding sample) based on quintiles. This made very little difference, with our conclusions entirely unaffected, and hence no results from this model are reported here.

Overall, the results presented here are disappointing, and indicate that:

(i) only two of the three ability measures (SDM and NART25) are significantly associated with wages (at least once age is controlled for);
(ii) cognitive ability explains just 1.3% of the variance in wages once demographic and human capital variables are controlled for, and as little as 0.5% after conditioning on job characteristics (such as occupation and industry); and
(iii) the estimated impacts of both the SDM and NART25 on wages are relatively small, with a one standard deviation in test scores associated with less than a 3% wage differential, conditional on personal and job-related characteristics.

Nevertheless, these conclusions are broadly consistent with findings from at least some previous international research and hence should not be entirely unexpected. Cawley, Heckman and Vytlacil (2001), for example, analysed data from the US National Longitudinal Study of Youth, which includes cognitive ability measures that are far superior to those included in the HILDA Survey, and concluded that the fraction of wage variance explained by cognitive ability was modest.[12] In their case, measured cognitive ability explained between 14% and 19% of wage variance in the absence of most controls. But once education and work experience were controlled for, this fell to between 0.7% and 2.7%. Our most comparable estimate, of 1.3% (from the row 4 specification), falls within this range.

One point of departure is the separability of the cognitive ability measures from education. Cawley et al. conclude that the correlation is so high that the two are inseparable. In contrast, the inclusion or exclusion of cognitive ability has relatively little effect on the estimated returns to education in the wave 12 HILDA Survey data. That said, the estimated penalty to low levels of education (not completing at least Year 11 of high school) is much smaller when conditioning on cognitive ability.

[12] Cawley et al. (2001) analysed data from the 1979 cohort of the NLSY which followed a sample of young people (aged between 13 and 20 in 1978) in the US, and who were administered the Armed Services Vocational Aptitude tests, a battery of 10 intelligence tests, in the second wave of the study.

Finally, and in line with other recent research (e.g., Heineck & Anger 2010), the estimated returns to non-cognitive skills are not much affected by the inclusion of measures of cognitive ability. This is reflected both in the results of the most complete specification reported in Table 11 (row 6), which includes the two achievement motivation scales, and in further analyses where I merged in data from wave 9 of the survey when data were collected, as part of the SCQ, on personality.

Described in more detail in Wooden (2012), a trait descriptive adjectives approach was used in both waves 5 and 9 to measure the Big Five personality traits – emotional stability, extroversion, openness to experience, agreeableness and conscientiousness. Multi-item scales measuring each of these dimensions are provided as part of the data set. These measures (from wave 9) were thus included in the wage specification (in the place of the two achievement motivation scales).[13] Only agreeableness and conscientiousness were found to exhibit statistically significant associations with hourly wages, with more agreeable workers facing a wage penalty while, as we would expect, more conscientious workers receiving a premium. The addition of the three cognitive ability measures, however, again had little effect on the estimated returns to these non-cognitive attributes.[14]

[13] This required sample members being observed in both wave 9 and wave 12, leading to a large drop in the sample size available for analysis (n=4052)

[14] The estimated parameter on the agreeableness scale moves closer to zero, and loses statistical significance but the actual magnitude of this change was relatively small (the estimated coefficient changed from -.015 to -.012). The coefficient on the conscientiousness scale is entirely unaffected (.024 in both specifications), while the coefficients on the other three dimensions remain small and insignificant.

Conclusions

The inclusion of measures of cognitive ability in the HILDA Survey was approached with a good deal of intrepidation, in part because of concerns about adverse reactions from sample members. It does appear, however, that such concerns were largely unfounded, with rates of task participation extremely high (over 95%), and feedback from interviewers that the large majority of respondents found the tasks a welcome change from the usual 'boring' questions that are asked very year.

While rates of participation were very high, non-participants were not a random group. Users of these variables may therefore need to give consideration to the creation of weights that adjust for non-random selection, both into face-to-face interviews (for the SDM and NART25 tasks) and into consenting to task participation.

Task administration seems to have been relatively straightforward, but nevertheless the fact that the tasks were not administered in a classroom-type setting could mean task scores were affected by outside influences. In particular, I find evidence of scores being significantly affected by the presence of someone else during the interview, the location of the interview, whether there was any significant distraction or disturbance while the task was being administered and, in the case of the BDS task, the interview mode. While the effects of each of these outside influences could be judged to be quite small, it may be wise to take them into account when working with the data.

Task scores can also be affected by the interviewer. This is especially problematic for the NART25, where interviewer judgment is required in determining whether a response constitutes a correct answer. Dealing with this, however, is relatively straightforward, provided users are not also interested in location-specific variations in task performance – interviewer effects will be correlated with location effects.

Task performance might also be affected by motivation, and a priori this might be expected to be most problematic for the SDM task where there is a time limit on task completion. The evidence does indeed suggest task performance is correlated with measures of achievement motivation, but is also more consistent with causation running from ability to motivation rather than the reverse.

Finally, I presented summary results from a simple regression-based examination of the association between the cognitive ability measures and hourly wages (among working-age employees). The results were rather disappointing. While two of the three measures were significantly associated with wages in multivariate models, the overall contribution of these measures to explaining the variation in wages was very small. More specifically, they appear to add very little to a model that already includes traditional human capital variables, such as measures of educational attainment, and nor is there much evidence that the estimated rates of return to education or to observed non-cognitive skills (such as motivation or personality) are much affected by the omission or inclusion of the cognitive ability measures.

References

Cattell, R.B. (1971). *Abilities: Their Structure, Growth, and Action*. New York: Houghton Mifflin.

Cawley, J., Heckman, J. & Vytlacil, E. (2001). Three observations on wages and measured cognitive ability. *Labour Economics* 8(4), 419-442.

Heineck, G. & Anger, S. (2010). Then returns to cognitive abilities and personality traits in Germany. *Labour Economics* 17(3), 535-546.

Kiely, K.M., Luscz, M.A., Piguet, O., Christensen, H., Bennett, H. & Anstey, K.J. (2011). Functional equivalence of the National Adult Reading Test (NART) and Schonell reading tests and NART norms on the Dynamic Analyses to Optimise Ageing (DYNOPTA) project. *Journal of Clinical and Experimental Neuropsychology* 33(4), 410-421.

Lachman, M.E. & Tun, P.A. (2008). Cognitive testing in large-scale surveys: assessment by telephone. In S. Hofer and D. Alwin (eds), *Handbook on Cognitive Aging: Interdisciplinary Perspectives* (pp. 506-523). Sage: Thousand Oaks (CA).

Lang, J.W.B. & Fries, S. (2006). A revised 10-item version of the achievement motives scale: psychometric properties in German-speaking samples. *European Journal of Psychological Assessment* 22(3), 216-224.

Nelson, H.E. (1982). *National Adult Reading Test (NART) Test Manual*. NFER-Nelson: Windsor (UK).

Smith, A. (2007). *Symbol Digits Modalities Test: Manual* (10^{th} printing). Western Psychological Services: Los Angeles.

Strauss, E., Sherman, E.M.S. & Spreen, O. (2006). *A Compendium of Neurospsychological Tests: Administration, Norms, and Commentary* (3rd ed.). Oxford University Press: Oxford.

Wooden, M. (2012). The stability of personality traits. In R. Wilkins and D. Warren, *Families, Incomes and Jobs, Volume 7: A Statistical Report on Waves 1 to 9 of the Household, Income and Labour Dynamics in Australia Survey*. Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Wooden, M., Mackinnon, A., Rodgers, B. & Windsor, T. (2012). The development of cognitive ability measures in the HILDA Survey (HILDA Survey Discussion Paper no. 1/12). Melbourne Institute of Applied Economic and Social Research, University of Melbourne. [http://www.melbourneinstitute.com/downloads/hilda/Bibliography/HILDA_Technical_Papers/htec112.pdf]

Appendix A: Paper Representation of CAPI Questionnaire Script, Cognitive Ability Tasks, Wave 12 Main Survey

Cognitive Ability Tasks (CATs)

INTRODUCTION TO BE READ OUT:

A special feature of the interview this year is the next section. It comprises three short exercises that involve you remembering and making judgement about words, symbols and numbers.

N10 I am going to read out some lists of numbers, and I want you to repeat the numbers back to me in the reverse order from which I said them. So if I said "3, 8", you would say "8, 3". Do you understand? Note that I cannot repeat the numbers after I have said them once.

If respondent does not understand, repeat the instructions.

The sets will get larger as we go. And it may help if you close your eyes to help you concentrate.

INTERVIEWER TO RECORD: IS IT OKAY TO START THE TASK?

Yes, start task..1 ➔N11
No, cannot understand instructions............................2 ➔N13
No, refused...3 ➔N13

When the respondent gets one trial correct at a "level" move on to the first trial at the next level. If the first trial is incorrect, administer the second trial. If both responses at the same level are incorrect, the test is discontinued.

Read in monotone, 1 second per number. Drop your voice on the last digit to indicate it is time to respond.

If participant immediately self-corrects, do not count as an error.

If the participant asks for repetition, say: "I'm sorry, I can't repeat items."

N11 READ OUT: **OK, I'll start now.**

	READ OUT	*CORRECT ANSWER*	*CORRECT*	*INCORRECT*
1a	**2, 4**	(4, 2)	1	2
1b	**5, 7**	(7, 5)	1	2
2a	**6, 2, 9**	(9, 2, 6)	1	2
2b	**4, 1, 5**	(5, 1, 4)	1	2
3a	**3, 2, 7, 9**	(9, 7, 2, 3)	1	2
3b	**4, 9, 6, 8**	(8, 6, 9, 4)	1	2
4a	**1, 5, 2, 8, 6**	(6, 8, 2, 5, 1)	1	2
4b	**6, 1, 8, 4, 3**	(3, 4, 8, 1, 6)	1	2
5a	**5, 3, 9, 4, 1, 8**	(8, 1, 4, 9, 3, 5)	1	2
5b	**7, 2, 4, 8, 5, 6**	(6, 5, 8, 4, 2, 7)	1	2
6a	**8, 1, 2, 9, 3, 6, 5**	(5, 6, 3, 9, 2, 1, 8)	1	2
6b	**4, 7, 3, 9, 1, 2, 8**	(8, 2, 1, 9, 3, 7, 4)	1	2
7a	**9, 4, 3, 7, 6, 2, 5, 8**	(8, 5, 2, 6, 7, 3, 4, 9)	1	2
7b	**7, 2, 8, 1, 9, 6, 5, 3**	(3, 5, 6, 9, 1, 8, 2, 7)	1	2

Once both trials at the same level are incorrect, say: **Ok, that's all of those we need to do.**

N12a INTERVIEWER RECORD: WAS THE TASK COMPLETED IN FULL?

Yes .... 1 ➔N12c
No.. 1 ➔N12b

N12b INTERVIEWER RECORD: WHY DID THE TASK HAVE TO BE CUT SHORT? *MULTI RESP*

Excessive distraction 1
Physical disability made completion impossible 2
Inability to understand the instructions 3
Extreme anxiety or discomfort 4
Refused to continue / doesn't want to do test 5
English language problems 6
Other (please specify) 7

N12c INTERVIEWER RECORD: WAS ANYONE ELSE PRESENT DURING THIS TASK? IF SO, WHO? *MULTI RESP*

Yes, another sample member .... 1 ➔N12d
Yes, child / children under 15 .... 2 ➔N12d
Yes, non-household member .... 3 ➔N12d
No, no one .... 4 ➔N12e

N12d INTERVIEWER RECORD: DID THIS PERSON HELP OR ASSIST THE RESPONDENT IN COMPLETING THE TASK?

Yes 1
No.. 2

N12e INTERVIEWER RECORD: WAS PERFORMANCE ON THIS TASK ADVERSELY AFFECTED BY ANY SIGNIFICANT DISTRACTION OR DISTURBANCE?

Yes 1
No.. 2

IF CONDUCTING INTERVIEW BY TELEPHONE, GO TO T1. OTHERWISE CONTINUE

N13 *Have Symbols Booklet to hand to respondent and stop-watch ready.*

This next exercise involves matching numbers to symbols.

(Hand self-complete card to participant)

Please look at the key at the top of the page. The symbol in the top row matches the number in the box below it.

Now look at the next line of boxes (*interviewer point to the line of boxes***). Notice that the boxes below the symbols are empty. Your task is to fill each empty box with the number that matches the symbol using the key at the top of the page. Is that clear?**

(If respondent requires further instructions say): Please look again at the key on top of the page, each of these symbols in the top row has a matching number. Your task is to fill in the blank boxes underneath each symbol (point) using the key at the top of the page to match the number? Is this clear?

Let's have a go at the first symbol. Looking at the key, you will see that number 1 goes in the first box, so write the number 1 in the first box. Now what number should you put in the second box? (*Number 5***) That's right. So write the number 5 in the second box. What number goes in the third box? (***Number 2***) Two, right.**

For practice, fill in the remaining boxes and stop at the double line.

Interviewer check practice boxes. Any errors made in these practice responses should be immediately pointed out. If needed, you will need to explain the task again.

Now when I say "Go!" write the numbers just like you have been doing until I say "Stop!", starting from here (*interviewer point to the first box after the double line***). When you come to the end of the first line, go quickly to the next line without stopping. If you make a mistake, just write the correct answer over your mistake. Don't skip any boxes and work as quickly as you can.**

Any questions?

INTERVIEWER TO RECORD: IS IT OKAY TO START THE TASK?

Yes, start task .... 1➔N14
No, cannot understand instructions........................2➔T1
No, refused...3➔T1

INTERVIEWER SAY: **OK, begin.**

N14 INTERVIEWER: AFTER THE 90 SECONDS IS UP, TALLY RESPONSES AND RECORD:

Correct responses ☐

N15a INTERVIEWER RECORD: WAS THE TASK COMPLETED IN FULL?

Yes ....1 ➔ N15c
No...2 ➔ N15b

N15b INTERVIEWER RECORD: WHY DID THE TASK HAVE TO BE CUT SHORT?
MULTI RESP

Excessive distraction 1
Physical disability made completion impossible 2
Inability to understand the instructions 3
Extreme anxiety or discomfort 4
Refused to continue / doesn't want to do test 5
English language problems 6
Other (please specify) 7

N15c INTERVIEWER RECORD: WAS ANYONE ELSE PRESENT DURING THIS TASK? IF SO, WHO? *MULTI RESP*

Yes, another sample member .... 1 ➔N15d
Yes, child / children under 15 .... 2 ➔N15d
Yes, non-household member .... 3 ➔N15d
No, no one .... 4 ➔N15e

N15d INTERVIEWER RECORD: DID THIS PERSON HELP OR ASSIST THE RESPONDENT IN COMPLETING THE TASK?

Yes 1
No 2

N15e INTERVIEWER RECORD: WAS PERFORMANCE ON THIS TASK ADVERSELY AFFECTED BY ANY SIGNIFICANT DISTRACTION OR DISTURBANCE?

Yes 1
No 2

N16 INTERVIEWER, INDICATE SHOWCARD N16

INTERVIEWER READ OUT: **I now want you to read slowly down this list of words. Start here** (*indicate first word)* **and read the word out loud. After each word please wait until I say "next" before reading out the next word. I must warn you that there are many words that you probably won't recognise; in fact <u>most</u> people don't know them, so just have a guess at these.**

INTERVIEWER TO RECORD: IS IT OKAY TO START THE TASK?

Yes, start task... 1➔N17
No, cannot understand instructions........................ 2➔T1
No, refused... 3➔T1

N17 Ok, go ahead with the first word.

		Acceptable pronunciations	*CORRECT*	*INCORRECT*
1	**CHORD**	kord	1	2
2	**AISLE**	ile	1	2
3	**DEBT**	det	1	2
4	**NAIVE**	ny-eev	1	2
5	**BOUQUET**	boo-kay, boe-kay	1	2
6	**PLACEBO**	ple-see-bo	1	2
7	**SUBTLE**	sut-l	1	2
8	**GOUGE**	gowj {"ow" as in "owl"}	1	2
9	**HIATUS**	hy-ay-tiss	1	2
10	**HEIR**	air	1	2
11	**EQUIVOCAL**	e-kwiv-e-kl, i-kwiv-e-kl, ee-kwiv-e-kl	1	2
12	**RAREFY**	rare-i-fy	1	2
13	**FACADE**	fa-sard, fassard	1	2
14	**ZEALOT**	zel-it	1	2
15	**SUPERFLUOUS**	sa-purf-loo-ess	1	2
16	**CELLIST**	chel-ist	1	2
17	**QUADRUPED**	kwod-roo-ped	1	2
18	**LEVIATHAN**	le-vy-e-then	1	2
19	**ABSTEMIOUS**	ab-stee-mee-us	1	2
20	**BEATIFY**	bee-at-i-fy	1	2
21	**SIDEREAL**	sy-deer-ee-el	1	2
22	**GAUCHE**	goe-sh	1	2
23	**DETENTE**	day-tont	1	2
24	**SYNCOPE**	sink-e-pee	1	2
25	**DEMESNE**	di-mayn, di-meen	1	2

N18a INTERVIEWER RECORD: WAS THE TASK COMPLETED IN FULL?

Yes .... 1 ➔N18c
No... 2 ➔N18b

N18b INTERVIEWER RECORD: WHY DID THE TASK HAVE TO BE CUT SHORT?
MULTI RESP

Excessive distraction 1
Physical disability made completion impossible 2
Inability to understand the instructions 3
Extreme anxiety or discomfort 4
Refused to continue / doesn't want to do test 5
English language problems 6
Other (please specify) 7

N18c INTERVIEWER RECORD: WAS ANYONE ELSE PRESENT DURING THIS TASK? IF SO, WHO? *MULTI RESP*

Yes, another sample member .... 1 ➔N18d
Yes, child / children under 15 .... 2 ➔N18d
Yes, non-household member .... 3 ➔N18d
No, no one .... 3 ➔N18e

N18d INTERVIEWER RECORD: DID THIS PERSON HELP OR ASSIST THE RESPONDENT IN COMPLETING THE TASK?

Yes 1
No 2

N18e INTERVIEWER RECORD: WAS PERFORMANCE ON THIS TASK ADVERSELY AFFECTED BY ANY SIGNIFICANT DISTRACTION OR DISTURBANCE?

Yes 1
No 2

Appendix B: Regression Results

Table B1: Dependent variables = Whether participated in task (and provided a valid score)
Estimation method = Logistic regression

	BDS		*SDM*		*NART25*	
	A	*B*	*A*	*B*	*A*	*B*
Female	.077	.104	.143*	.228**	.232**	.283**
	(.070)	(.078)	(.072)	(.083)	(.070)	(.075)
Age group (ref. cat. = 25-29)						
15-19	.433*	.654**	.400	.877**	.616**	.688**
	(.198)	(.220)	(.220)	(.253)	(.177)	(.191)
20-24	.207	.131	.068	.024	.036	-.124
	(.189)	(.202)	(.211)	(.226)	(.167)	(.178)
30-34	-.021	.037	-.215	-.097	.120	.105
	(.191)	(.208)	(.210)	(.233)	(.183)	(.196)
35-39	-.098	-.032	-.373	-.236	-.007	-.071
	(.186)	(.203)	(.201)	(.224)	(.176)	(.188)
40-44	-.117	-.048	-.384	-.212	.184	.202
	(.181)	(.198)	(.197)	(.218)	(.177)	(.190)
45-49	-.281	-.324	-.236	-.182	.127	.051
	(.176)	(.192)	(.200)	(.220)	(.173)	(.185)
50-54	-.359*	-.212	-.390*	-.223	.213	.207
	(.173)	(.191)	(.1940	(.216)	(.175)	(.189)
55-59	-.200	-.231	-.206	-.125	.114	.049
	(.182)	(.199)	(.203)	(.225)	(.175)	(.188)
60-64	-.316	-.231	-.462*	-.354	.105	.055
	(.181)	(.198)	(.198)	(.220)	(.178)	(.193)
65-69	-.275	-.357	-.247	-.187	.196	.135
	(.188)	(.206)	(.210)	(.233)	(.184)	(.199)
70-74	-.545**	-.497*	-.569**	-.426	.262	.289
	(.189)	(.214)	(.207)	(.236)	(.194)	(.215)
75-79	-.459*	-.467*	-.890**	-1.017**	.236	.202
	(.208)	(.233)	(.211)	(.236)	(.210)	(.228)
80+	-.749**	-1.036**	-1.393**	-1.768**	-.080	-.311
	(.186)	(.207)	(.190)	(.210)	(.183)	(.197)
Long-term health condition (ref cat = No condition)						
Mild	-.372**	-.253	-.338**	-.316*	-.277*	-.300*
	(.116)	(.135)	(.119)	(.142)	(.119)	(.132)
Moderate	-.333**	-.524**	-.411**	-.782**	-.405**	-.596**
	(.091)	(.105)	(.092)	(.110)	(.090)	(.100)
Severe	-1.209**	-1.294**	-1.692**	-1.966**	-1.520**	-1.550**
	(.140)	(.166)	(.133)	(.166)	(.138)	(.158)
Origin (ref. cat = Australian-born, non-indigenous)						
Indigenous Australian	-.237	-.411	-.344	-.241	-.706**	-.726**
	(.199)	(.223)	(.202)	(.246)	(.158)	(.174)
OS born: English language	.244	.105	.104	.080	.323*	.235
	(.126)	(.142)	(.123)	(.144)	(.136)	(.147)
OS born: Other language	-.437**	-.537**	-.501**	-.525**	-.478**	-.565**
	(.148)	(.163)	(.150)	(.167)	(.146)	(.158)

Table B1 (cont'd)

	BDS		*SDM*		*NART25*	
	A	*B*	*A*	*B*	*A*	*B*
English speaking ability (ref. cat. = Speaks only English at home)						
Very well	-.154	.094	-.289	-.080	-.443**	-.261
	(.159)	(.176)	(.160)	(.181)	(.151)	(.166)
Well	-.858**	-.953**	-.757**	-.774**	-1.416**	-1.479**
	(.180)	(.203)	(.186)	(.211)	(.171)	(.191)
Not well / not at all	-1.850**	-2.290**	-1.792**	-2.127**	-2.728**	-3.040**
	(.192)	(.233)	(.198)	(.237)	(.194)	(.230)
Highest education level (ref. cat. = Year 11 or less)						
Higher degree / qual.	.659**	.885**	.706**	.966**	1.604**	1.792**
	(.143)	(.161)	(.145)	(.169)	(.171)	(.189)
Bachelor degree	.660**	.755**	.907**	1.065**	1.433**	1.583**
	(.126)	(.141)	(.138)	(.156)	(.139)	(.154)
Diploma	.625**	.733**	.909**	1.073**	1.270**	1.334**
	(.144)	(.162)	(.159)	(.183)	(.160)	(.171)
Certificate III / IV	.565**	.705**	.628**	.853**	.820**	.894**
	(.101)	(.114)	(.103)	(.120)	(.096)	(.104)
Year 12	.334**	.460**	.470**	.630**	1.037**	1.158**
	(.110)	(.124)	(.116)	(.133)	(.115)	(.125)
Others present during	-.623**	-.659**	-.434**	-.466**	-.533**	-.557**
interview	(.071)	(.083)	(.072)	(.086)	(.070)	(.070)
Interview mode / location (ref. cat = face-to-face at home)						
Face-to-face: outside home	-.354	-.663**	-.300	-.782**	-.607**	-.659**
	(.218)	(.264)	(.220)	(.272)	(.190)	(.077)
Face-to-face: workplace	-.374	.317	-.174	.398	.125	.041
	(.428)	(.596)	(.468)	(.695)	(.524)	(.554)
Face-to-face: other place	-.639**	-.855**	-.816**	-1.163**	-.992**	-1.028**
	(.225)	(.245)	(.212)	(.233)	(.191)	(.205)
Phone	-.597**	-1.537**				
	(.135)	(.253)				
Constant	3.298**	1.653**	3.291**	1.750**	2.537	1.733
	(.158)	(.282)	(.174)	(.312)	(.145)	(.346)
Cox and Snell R-squared	.045	.131	.060	.152	.080	.137
Nagelkerke R-squared	.128	.368	.164	.417	.206	.354
Chi-squared	801.5**	2418.1	991.1**	2641.9**	1341.8**	2376.1**
N	17237	17237	16070	16070	16069	16069

Notes: Model B includes 176 interviewer dummies; Model A does not.
Figures in parentheses are standard errors.
* and ** denote statistical significance at the 95% and 99% confidence levels, respectively.

Table B2: Dependent variables = Cognitive ability task scores
Estimation method = Ordinary least squares

	BDS		*SDM*		*NART25*	
Explanatory variable	*Coeff.*	*Std. err.*	*Coeff.*	*Std. err.*	*Coeff.*	*Std. err.*
Female	.031	.022	3.241**	.167	.219*	.077
Age group (ref. cat. = 25-29)						
15-19	.085	.053	1.257**	.397	-.213	.184
20-24	.003	.050	.066	.377	.050	.175
30-34	.042	.052	-.959*	.396	.436*	.183
35-39	.063	.052	-2.484**	.391	1.138**	.180
40-44	.019	.051	-3.298**	.386	1.327**	.178
45-49	.065	.052	-5.480**	.390	2.042**	.180
50-54	-.025	.052	-6.691**	.394	2.133**	.181
55-59	-.037	.054	-8.196**	.407	2.640**	.188
60-64	-.032	.057	-10.722**	.432	3.234**	.199
65-69	-.138*	.060	-13.839**	.450	3.281**	.208
70-74	-.179**	.067	-17.428**	.494	3.579**	.227
75-79	-.302**	.076	-20.916**	.565	3.646**	.256
80+	-.524**	.072	-25.312**	.541	3.340**	.242
Long-term health condition (ref cat = No condition)						
Mild	-.108**	.041	-1.281**	.297	-.143	.136
Moderate	-.130**	.033	-3.517**	.242	-.598**	.112
Severe	-.497**	.080	-7.516**	.633	-1.962**	.286
Origin (ref. cat = Australian-born, non-indigenous)						
Indigenous Australian	-.389**	.067	-3.484**	.511	-2.148**	.241
OS born: English language	.016	.035	.247	.263	.226	.120
OS born: Other language	-.163**	.052	-.017	.386	-2.334**	.177
English speaking ability (ref. cat. = Speaks only English at home)						
Very well	-.107*	.050	-1.021**	.375	-1.162**	.172
Well	-.486**	.079	-5.212**	.580	-5.131**	.274
Not well / not at all	-.553**	.122	-7.746**	.887	-7.400**	.467
Highest education level (ref. cat. = Year 11 or less)						
Higher degree / qual.	.769**	.042	6.663**	.314	6.199**	.144
Bachelor degree	.773**	.038	6.349**	.284	5.472**	.131
Diploma	.425**	.043	4.773**	.323	3.289**	.148
Certificate III / IV	.114**	.033	1.676**	.246	1.187**	.113
Year 12	.422**	.035	4.220**	.265	3.189**	.122
Interviewer observations						
Others present during task	-.079**	.025	-.375*	.185	-.479**	.085
Task adversely affected	.018	.042	-.825*	.410	-.164	.228
Interview mode / location (ref. cat = face-to-face at home)						
Face-to-face: outside home	.120	.080	-1.092	.577	-.764**	.270
Face-to-face: workplace	.124	.133	1.894*	.951	.001	.433
Face-to-face: other place	.067	.089	1.169	.644	-.370	.301
Phone	.136**	.042				
Achievement motivation						
Fear of failure	-.044**	.009	-.504**	.066	-.118**	.035
Hopes for success	.046**	.012	.340**	.085	.168**	.039
SCQ not returned	-.099	.084	-2.969**	.636	-.333	.293
Constant	3.648**	.089	51.997**	.658	8.802**	.323

Table B2 (cont'd)

	BDS	*SDM*	*NART25*
Adjusted R-squared	.084	.421	.306
F	40.528**	298.17**	180.33**
N	16039	14734	14622

Notes: * and ** denote statistical significance at the 95% and 99% confidence levels, respectively.